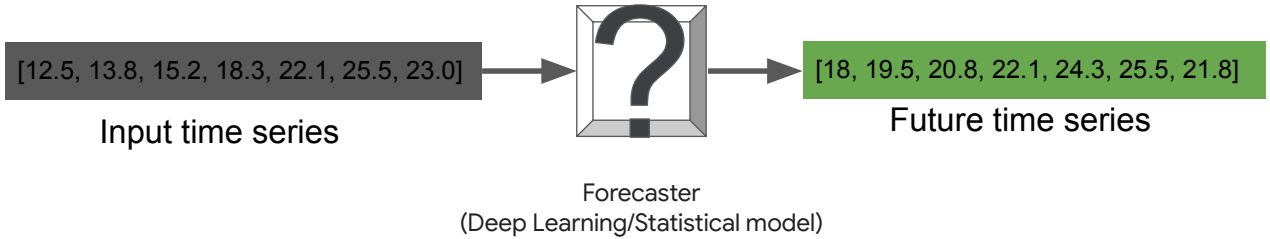
Google Cloud

TFRBench: A Reasoning Benchmark for Evaluating Forecasting Systems

Md Atik Ahamed, Mihir Parmar, Palash Goyal, Yiwen Song, Long T. Le, Qiang Cheng, Chun-Liang Li, Hamid Palangi, Jinsung Yoon, Tomas Pfister

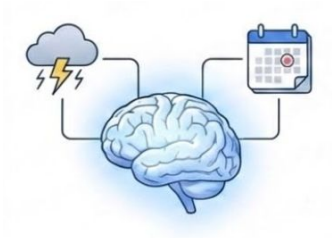
Time Series Forecasting in General



But what about context?
And what about reasoning?

The Gap in Reliability and Explainability

The “Why” Matters



A model might correctly predict a sudden spike in electricity demand, but it is much more valuable if it understands the cause, such as a specific weather event or a holiday.

The Challenge of Evaluation

Unlike numerical targets, there is no natural “ground truth” for the reasoning behind a forecast in existing benchmarks.

The Void

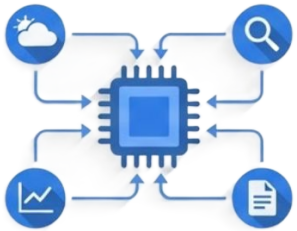
Currently, there are no standardized benchmarks that capture the reasoning quality behind a numerical forecast.

Introducing TFRBench

What is it?

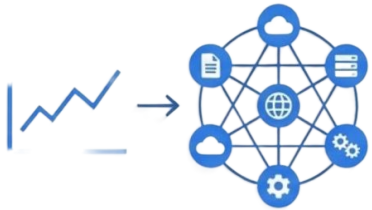
The first benchmark specifically designed to evaluate the reasoning capabilities of forecasting systems alongside their numerical accuracy.

Defining “Reasoning”



In this context, reasoning is a model's ability to analyze multiple data channels, identify interdependencies, and provide helpful directive for future forecasting.

The Paradigm Shift



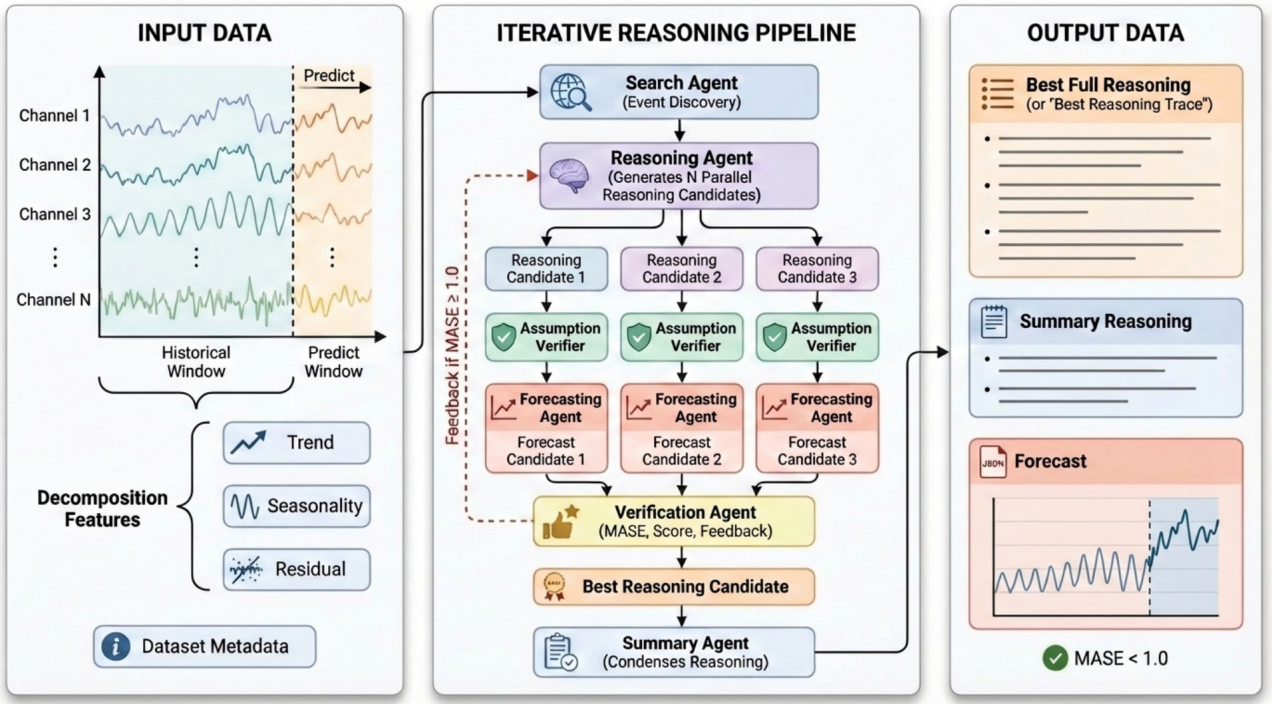
Moving from isolated numerical sequences to forecasts derived from explicit, verifiable causal logic.

The Scope of TFRBench

- **Broad Generalization:** Designed to test reasoning across varying data distributions.
- **Diverse Domains:** It spans 10 distinct datasets across 5 high-stakes domains.
- **Diverse temporal frequencies:** (Hourly to Monthly)
- **Univariate to Multivariate:** Expands from univariate (i.e., single channel) to multivariate (e.g., 5 channels)

Domain	Dataset	Source	Total Samples	Reference Samples	Freq.	Window (in-out)	Ch.	Series
Energy	Solar	GIFT-Eval	274	193	Daily	96-96	1	137
	Electricity	GIFT-Eval	1346	353	Daily	96-96	1	81
Sales	Car-parts	GIFT-Eval	1037	692	Monthly	25-25	1	1037
	Hierarchical Sales	GIFT-Eval	1370	1076	Daily	96-96	1	81
Web/CloudOps	Bitbrains - Fast Storage	GIFT-Eval	1153	678	Hourly	96-96	2	197
	Web Traffic	Monash archive	1214	710	Daily	96-96	1	162
Transportation	Traffic	Autoformer et al.	1104	449	Hourly	96-96	1	6
	NYC Taxi	nyc.gov	1000	391	Hourly	96-96	1	1
Economics/Finance	Amazon pricing	Yahoo Finance/nasdaq	512	328	Daily	14-7	5	1
	Apple pricing	Yahoo Finance/nasdaq	809	509	Daily	14-7	5	1

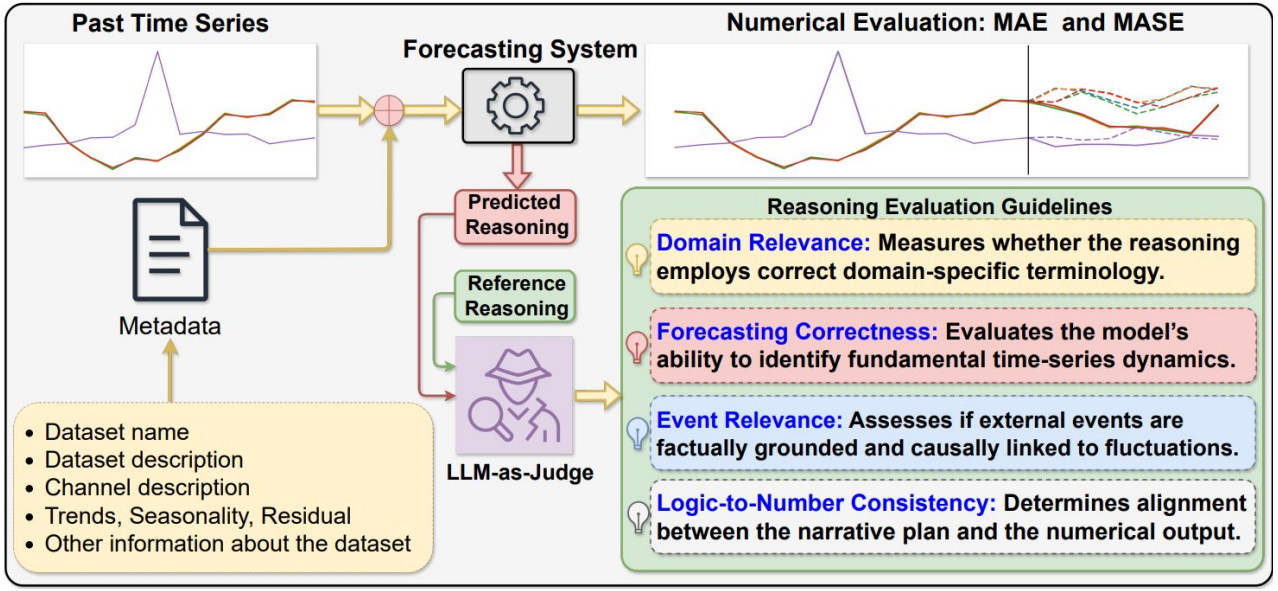
Overview of Data Creation System



Iterative "generate, verify, and refine" cycle

5 specialized agents: Search, Reasoning, Forecasting, Verification, Summary

Overview of the TFRBench Evaluation Pipeline



We employ an LLM-as-a-Judge to audit a candidate's reasoning trace against our Reference Reasoning.

The core focus of our framework is evaluating the quality of the reasoning.

Reasoning Evaluator (LLM-as-a-Judge)

Objective: Our goal is to evaluate the actual reasoning process behind forecasting. We use an LLM-as-a-Judge to audit candidates against "Reference Reasoning", verifying their logical, factual, and domain-specific accuracy.

Domain Relevance	Forecasting Correctness	Event Relevance	Logic-to-Number Consistency
1. **(Irrelevant/Wrong):** Wrong domain terminology. Logic makes no sense. 2. **(Generic):** Vague language without domain terms. 3. **(Acceptable):** Basic terms used correctly but lacks depth. 4. **(Good):** Specific terminology and standard forecasting logic used correctly. 5. **(Expert):** Deep domain expertise, precise jargon, matches Ground Truth logic.	1. **(Incorrect):** Completely misses trend or seasonality. 2. **(Weak):** Identifies trend but ignores obvious seasonality. 3. **(Average):** Captures main trend/seasonality but misses magnitude/shape. 4. **(Strong):** Correctly identifies trend, seasonality, and general shape. 5. **(Exact):** Perfectly captures trend, seasonality, and inflection points.	1. **(Hallucination):** Mentions non-existent or false events. 2. **(Irrelevant):** Real events but no logical connection. 3. **(Correlated):** Identifies event but explanation is vague. 4. **(Causal):** Clearly links event to specific data movement. 5. **(Aligned):** Detailed, accurate causal explanation of impact.	1. **(Contradiction):** Text and numbers are opposites. 2. **(Disconnected):** Text describes shapes not seen in numbers. 3. **(Weak Consistency):** Direction matches but magnitude is off. 4. **(Consistent):** Numbers generally follow the narrative. 5. **(Alignment):** Numerical forecast is a precise translation of the reasoning.

LLM-as-a-Judge Results

Dataset	Models	w/ Reasoning				Event Forecast + Reasoning			
		Dom.	Fcst.	Evt.	Logic	Dom.	Fcst.	Evt.	Logic
Solar Daily	Gemini-2.5-Pro	$3.026_{0.000}$	$2.316_{0.000}$	$2.036_{0.000}$	$4.528_{0.000}$	$3.803_{0.000}$	$2.731_{0.000}$	$2.207_{0.000}$	$4.207_{0.000}$
	Gemini-3-Pro	$4.699_{0.000}$	$3.212_{0.000}$	$2.508_{0.000}$	$4.881_{0.000}$	$4.119_{0.000}$	$2.699_{0.000}$	$2.368_{0.000}$	$4.523_{0.000}$
	Claude-Sonnet-4.5	$4.922_{0.000}$	$3.062_{0.000}$	$2.891_{0.000}$	$3.363_{0.000}$	$4.554_{0.000}$	$2.637_{0.000}$	$2.472_{0.000}$	$3.368_{0.000}$
Electricity	Gemini-2.5-Pro	$2.318_{0.013}$	$2.768_{0.020}$	$1.720_{0.022}$	$4.590_{0.015}$	$4.688_{0.000}$	$3.076_{0.000}$	$3.442_{0.000}$	$3.198_{0.000}$
	Gemini-3-Pro	$4.167_{0.000}$	$3.173_{0.000}$	$2.178_{0.000}$	$4.773_{0.000}$	$4.892_{0.000}$	$3.606_{0.000}$	$4.147_{0.000}$	$4.484_{0.000}$
	Claude-Sonnet-4.5	$3.810_{0.000}$	$2.054_{0.000}$	$2.093_{0.000}$	$3.167_{0.000}$	$4.609_{0.000}$	$1.971_{0.001}$	$2.885_{0.001}$	$2.203_{0.001}$
Car Parts	Gemini-2.5-Pro	$3.123_{0.000}$	$2.658_{0.000}$	$1.863_{0.000}$	$4.181_{0.000}$	$4.926_{0.000}$	$3.464_{0.000}$	$3.811_{0.000}$	$4.923_{0.000}$
	Gemini-3-Pro	$4.068_{0.000}$	$2.900_{0.000}$	$2.088_{0.000}$	$4.945_{0.000}$	$4.962_{0.000}$	$3.012_{0.000}$	$4.158_{0.001}$	$4.883_{0.000}$
	Claude-Sonnet-4.5	$4.120_{0.000}$	$2.675_{0.000}$	$2.022_{0.000}$	$3.384_{0.000}$	$4.939_{0.000}$	$3.402_{0.001}$	$4.084_{0.000}$	$4.402_{0.000}$
Hierarchical Sales	Gemini-2.5-Pro	$2.637_{0.014}$	$2.972_{0.007}$	$1.635_{0.010}$	$4.730_{0.015}$	$2.519_{0.000}$	$2.456_{0.000}$	$2.230_{0.000}$	$3.346_{0.000}$
	Gemini-3-Pro	$3.701_{0.000}$	$2.878_{0.000}$	$1.967_{0.000}$	$4.335_{0.000}$	$2.476_{0.000}$	$2.492_{0.000}$	$2.205_{0.000}$	$3.798_{0.000}$
	Claude-Sonnet-4.5	$3.757_{0.000}$	$2.613_{0.000}$	$2.046_{0.000}$	$3.651_{0.000}$	$2.198_{0.000}$	$2.019_{0.000}$	$1.947_{0.000}$	$2.836_{0.000}$
Bitbrains Fast Storage	Gemini-2.5-Pro	$3.807_{0.000}$	$2.565_{0.000}$	$2.944_{0.000}$	$3.845_{0.000}$	$3.802_{0.000}$	$1.313_{0.003}$	$2.652_{0.001}$	$1.324_{0.000}$
	Gemini-3-Pro	$4.531_{0.000}$	$3.451_{0.000}$	$3.249_{0.000}$	$4.879_{0.000}$	$4.746_{0.000}$	$2.963_{0.001}$	$2.704_{0.006}$	$4.174_{0.002}$
	Claude-Sonnet-4.5	$4.714_{0.000}$	$2.780_{0.000}$	$3.427_{0.001}$	$2.507_{0.000}$	$3.360_{0.000}$	$1.432_{0.000}$	$2.083_{0.000}$	$1.543_{0.000}$
Web Traffic	Gemini-2.5-Pro	$3.585_{0.000}$	$2.831_{0.000}$	$2.039_{0.000}$	$4.693_{0.000}$	$4.576_{0.000}$	$2.761_{0.000}$	$3.249_{0.000}$	$3.645_{0.000}$
	Gemini-3-Pro	$4.034_{0.000}$	$3.169_{0.000}$	$2.179_{0.000}$	$4.904_{0.000}$	$4.634_{0.000}$	$2.804_{0.000}$	$3.435_{0.000}$	$4.699_{0.000}$
	Claude-Sonnet-4.5	$3.928_{0.000}$	$2.325_{0.000}$	$2.120_{0.000}$	$3.727_{0.000}$	$4.276_{0.000}$	$2.097_{0.000}$	$2.669_{0.000}$	$2.637_{0.001}$
Traffic	Gemini-2.5-Pro	$2.762_{0.012}$	$1.946_{0.004}$	$1.896_{0.027}$	$4.584_{0.034}$	$3.643_{0.001}$	$3.143_{0.000}$	$2.203_{0.000}$	$4.777_{0.000}$
	Gemini-3-Pro	$3.793_{0.000}$	$2.483_{0.000}$	$2.249_{0.000}$	$4.842_{0.000}$	$2.604_{0.000}$	$2.664_{0.000}$	$2.022_{0.000}$	$3.935_{0.000}$
	Claude-Sonnet-4.5	$4.294_{0.000}$	$2.254_{0.000}$	$2.488_{0.000}$	$3.586_{0.000}$	$3.381_{0.000}$	$2.379_{0.000}$	$1.962_{0.000}$	$3.454_{0.000}$
Nyc Taxi	Gemini-2.5-Pro	$2.814_{0.018}$	$1.922_{0.010}$	$1.703_{0.018}$	$4.360_{0.040}$	$4.972_{0.000}$	$3.529_{0.001}$	$3.425_{0.000}$	$4.652_{0.000}$
	Gemini-3-Pro	$3.969_{0.000}$	$2.678_{0.000}$	$2.113_{0.000}$	$4.872_{0.000}$	$4.941_{0.000}$	$3.844_{0.000}$	$3.941_{0.000}$	$4.555_{0.000}$
	Claude-Sonnet-4.5	$4.532_{0.000}$	$2.297_{0.000}$	$2.363_{0.000}$	$2.767_{0.000}$	$4.890_{0.000}$	$2.719_{0.000}$	$3.156_{0.000}$	$3.097_{0.000}$
Amazon	Gemini-2.5-Pro	$2.262_{0.000}$	$1.488_{0.000}$	$1.637_{0.000}$	$3.213_{0.000}$	$4.966_{0.000}$	$2.596_{0.001}$	$3.617_{0.001}$	$4.640_{0.002}$
	Gemini-3-Pro	$2.936_{0.000}$	$1.829_{0.000}$	$1.784_{0.000}$	$3.976_{0.000}$	$4.997_{0.000}$	$3.433_{0.000}$	$4.125_{0.000}$	$4.857_{0.000}$
	Claude-Sonnet-4.5	$3.588_{0.000}$	$2.012_{0.000}$	$1.893_{0.000}$	$3.970_{0.000}$	$4.988_{0.000}$	$2.543_{0.000}$	$3.179_{0.001}$	$4.491_{0.000}$
Apple	Gemini-2.5-Pro	$2.217_{0.001}$	$1.557_{0.001}$	$1.695_{0.000}$	$3.380_{0.002}$	$4.986_{0.000}$	$2.556_{0.000}$	$3.648_{0.000}$	$4.642_{0.000}$
	Gemini-3-Pro	$3.000_{0.000}$	$1.842_{0.001}$	$1.800_{0.000}$	$3.821_{0.001}$	$5.000_{0.000}$	$3.327_{0.001}$	$4.313_{0.001}$	$4.826_{0.002}$
	Claude-Sonnet-4.5	$3.617_{0.000}$	$2.075_{0.000}$	$1.917_{0.000}$	$3.786_{0.000}$	$4.996_{0.000}$	$2.515_{0.000}$	$3.387_{0.000}$	$4.642_{0.001}$

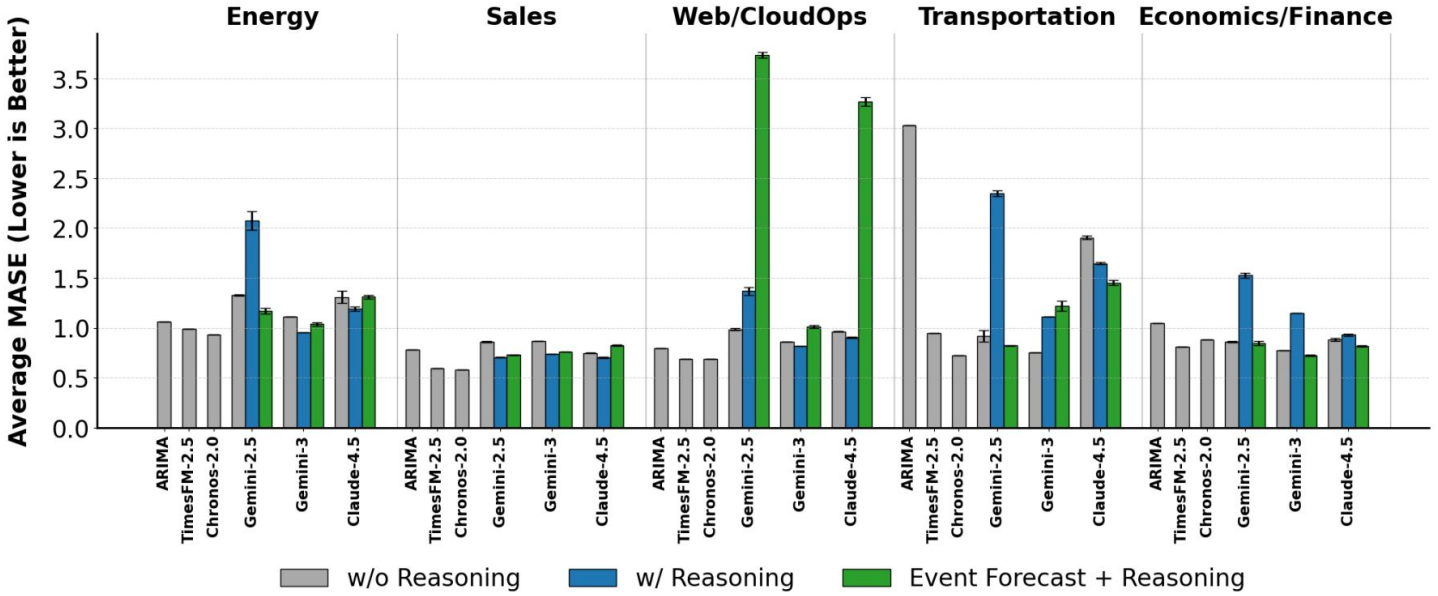
Majority of the models are not good at event relevance.

Though, they are good at logic to number conversion, their forecasting correctness are poor.

As forecasting correctness are poor, their reasoning is suffered from narrative bias.

This shows the risk of relying on regular LLM capabilities without the reasoning verification, TFRBench makes an effort to enable this.

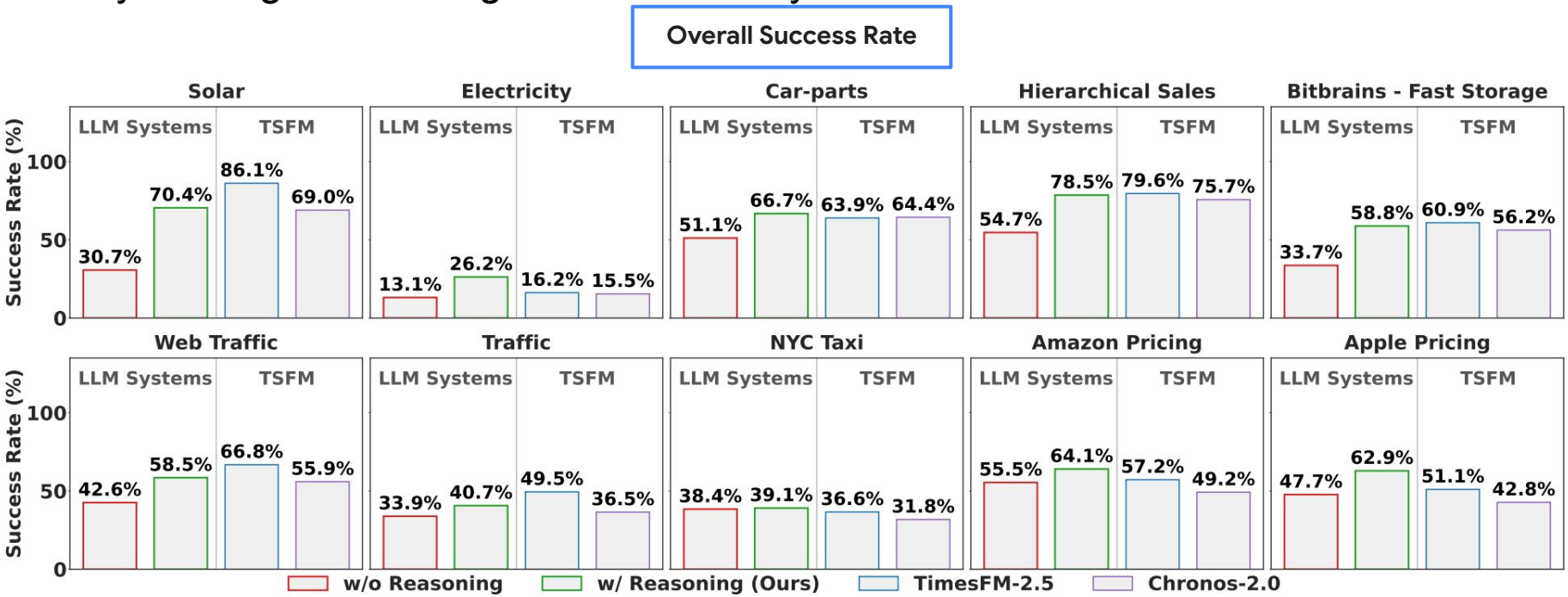
Avg. MASE performance by domains



Accessing context for past data (e.g., weather, holidays) allows models to reason better, which is highly effective in Energy and Transportation.

However, this strategy is not always beneficial. In Web/CloudOps, event-aware reasoning acts as a distractor.

Key Finding: Reasoning Drives Accuracy



Prompting LLMs with our generated reasoning traces significantly improves forecasting accuracy compared to direct numerical prediction.

LLMs are zero shot, while TimesFM-2.5 and Chronos-2.0 are pretrained with large numeric forecasting data.

Conclusion and Impact

A New Standard

TFRBench establishes a new benchmark for **interpretable**, **reasoning-based** evaluation in time-series forecasting.

Accountability

By requiring models to explain their predictions in high-stakes domains (Energy, Finance), we enhance human trust and accountability in automated decision-making.

Diagnostic Power

TFRBench serves as a vital diagnostic tool to identify biases and ensure forecasts are driven by factual logic, not statistical artifacts.